HW7

Zhenmin Hua

5905057247

Task2-1

Q: What is the size of the representation vector?
A: 384 (each)

Task2-2

Q: How many clusters are created for each algorithm?
What is the running time cost of each algorithm?
A: fastclustering: 61 clusters; 42.47 sec
Kmeans: **OOM** `IOPub data rate exceeded.`

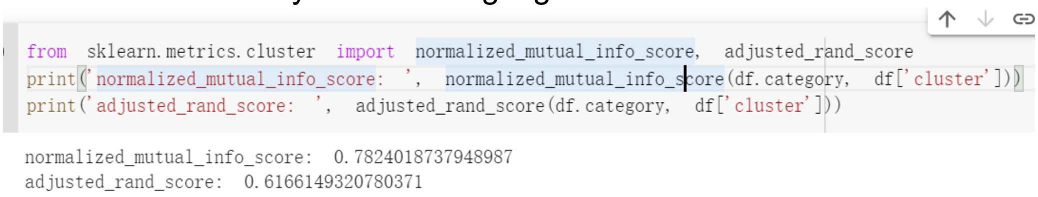
Task2-3

Q: What is the size of the representation vector after dimension reduction?
A: 128

Q: How many clusters are created for each algorithm?
What is the running time cost of each algorithm?
A: fastclustering: 94 clusters; 24.22 sec
Kmeans: **OOM** `IOPub data rate exceeded.`
(I tried to change the dimension of sentence representation, however, still OOM)

Task2-3

Q: What is the score of your clustering algorithm?

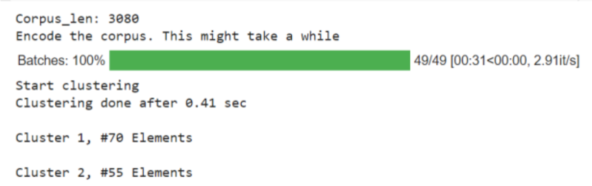


```
from sklearn.metrics.cluster import normalized_mutual_info_score, adjusted_rand_score
print('normalized_mutual_info_score: ', normalized_mutual_info_score(df.category, df['cluster']))
print('adjusted_rand_score: ', adjusted_rand_score(df.category, df['cluster']))
```

```
normalized_mutual_info_score:  0.7824018737948987
adjusted_rand_score:  0.6166149320780371
```

A:

Q: What is the running time cost of each algorithm?
A: I am using fast-clustering method. And the time cost is 0.41 sec.



```
Corpus_len: 3080
Encode the corpus. This might take a while
Batches: 100%|██████████| 49/49 [00:31<00:00, 2.91it/s]
Start clustering
Clustering done after 0.41 sec

Cluster 1, #70 Elements

Cluster 2, #55 Elements
```